

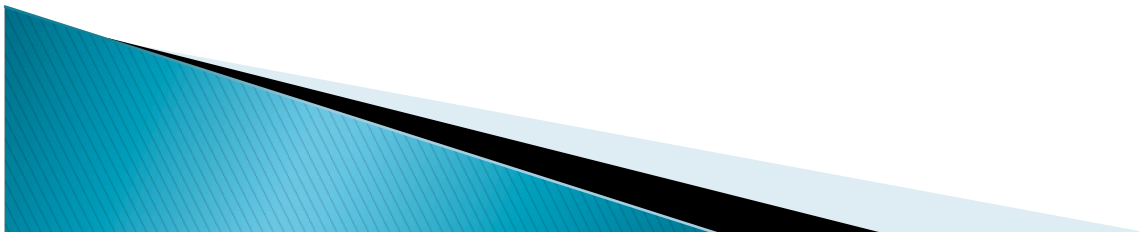


Common Data Challenges in the Great Survey Era

Alex Szalay
The Johns Hopkins University

Outline

- ▶ Today's challenges
- ▶ SDSS lessons
- ▶ The road ahead: NVO+VAO
- ▶ Trends: how long can this continue?



Survey Trends

CMB Surveys (pixels)

▶ 1990	COBE	1 000
▶ 2000	Boomerang	10,000
▶ 2002	CBI	50,000
▶ 2003	WMAP	1 Million
▶ 2008	Planck	10 Million

Angular Galaxy Surveys (obj)

• 1970	Lick	1M
• 1990	APM	2M
• 2005	SDSS	200M
• 2008	PS1	1000M
• 2010	VISTA	1000M
• 2014	LSST	3000M

Time Domain

- QUEST
- SDSS Extension survey
- Dark Energy Camera
- PanStarrs
- JDAM...
- LSST...

Galaxy Redshift Surveys (obj)

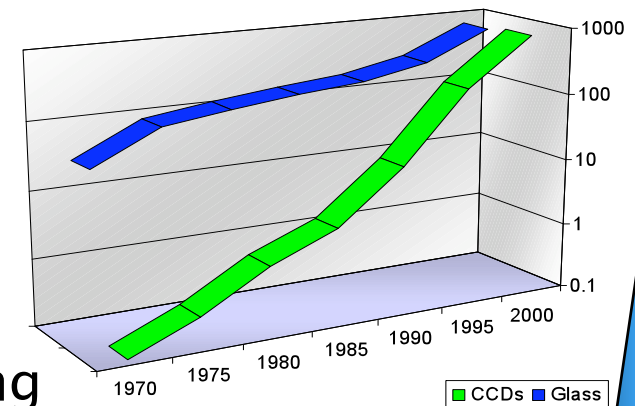
• 1986	CfA	3500
• 1996	LCRS	23000
• 2003	2dF	250000
• 2005	SDSS	750000

Petabytes/year by the end of the decade...



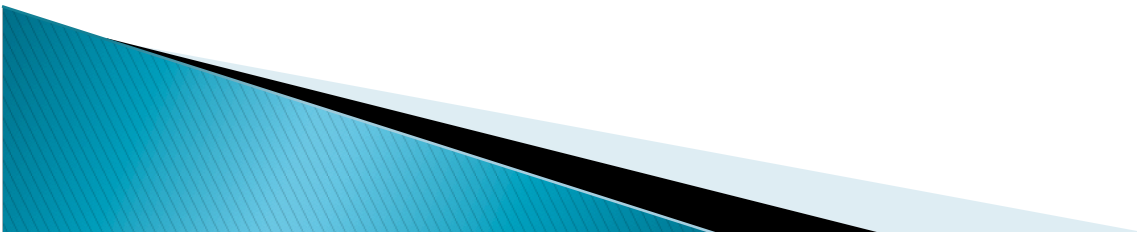
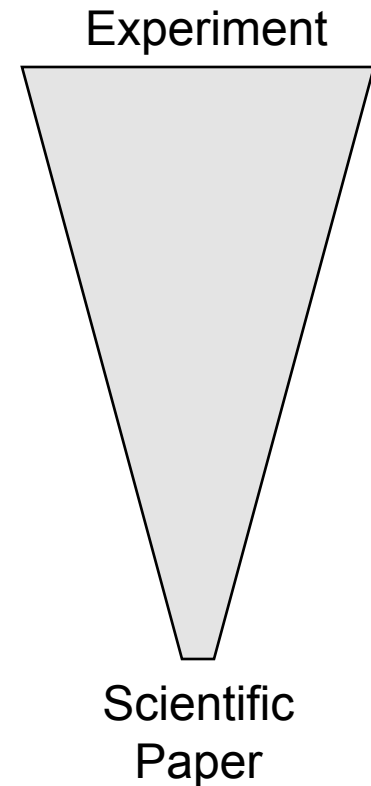
An Exponential World

- ▶ Scientific data doubles every year
 - caused by successive generations of inexpensive sensors + exponentially faster computing
- ▶ Changes the nature of scientific computing
- ▶ Cuts across disciplines (eScience)
- ▶ It becomes increasingly harder to extract knowledge
- ▶ 20% of the world's servers go into data centers by the "Big 5"
 - Google, Microsoft, Yahoo, Amazon, eBay
- ▶ So it is not only the scientific data!



The Data Explosion

- ▶ We see the **Industrial Revolution** in collecting scientific data
- ▶ Main Steps:
 - Acquire data (doubling)
 - Process/calibrate
 - Transform and load
 - Reorganize
 - Analyze/collaborate
 - Publish (~constant)



Technical Challenges

▶ **Data Access:**

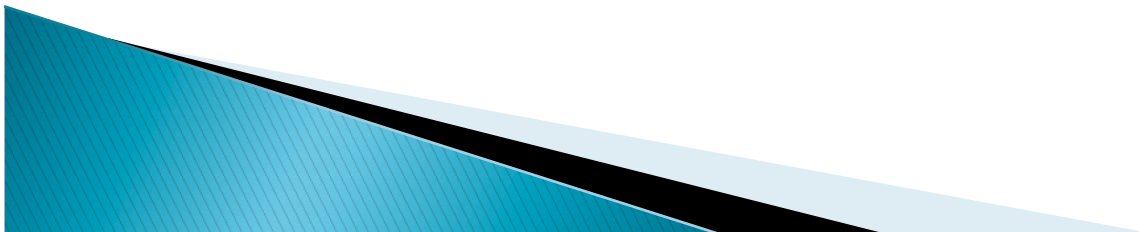
- Data sets have a power law distribution
- Move analysis to the data
- Locality is the key

▶ **Discovery:**

- Shannon \Leftrightarrow new dimensions
- Federation still requires data movement

▶ **Analysis:**

- Only max $N \log N$ algorithms possible



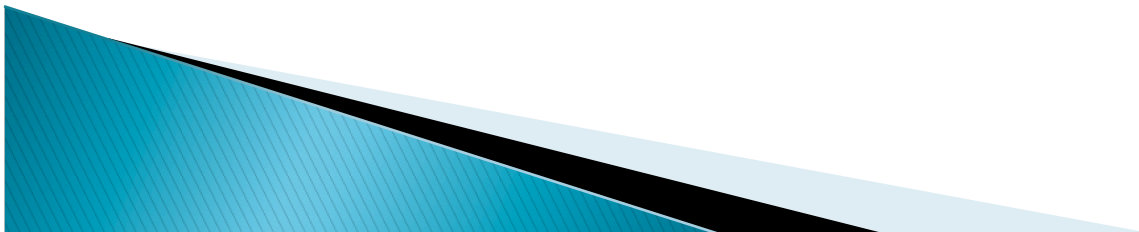
SDSS Now Finished!

- ▶ As of May 15, 2008 SDSS is officially complete
- ▶ Final data release (DR7) on Oct 31, 2008
- ▶ Final archiving of the data in progress
 - Paper archive at U. Chicago Library
 - Deep Digital Archive at JHU Library
 - CAS Mirrors at FNAL+JHU P&A
- ▶ Archive will contain >100TB
 - All raw data
 - All processed/calibrated data
 - All versions of the database (>18TB)
 - Full email archive and technical drawings
 - Full software code repository
 - Telescope sensor stream, IR fisheye camera, etc



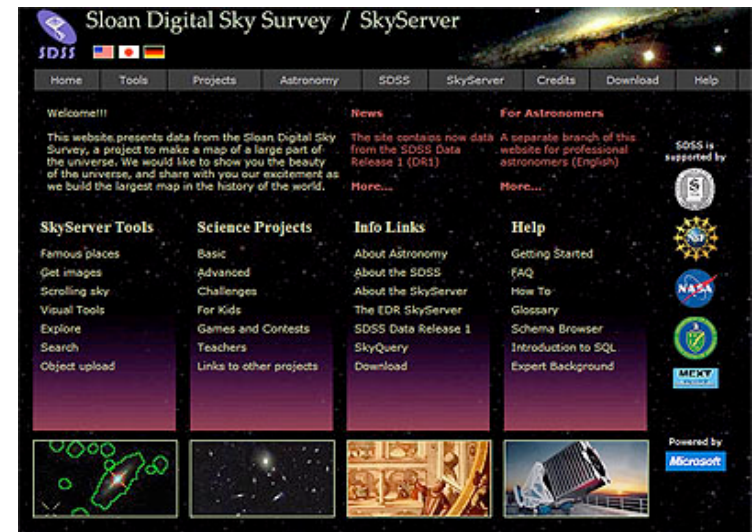
Capture Communications

- ▶ No 'Einstein letters' today... very little paper trail
- ▶ Proposals and papers archived
- ▶ Most large projects communicate through email exploders and phonecons
- ▶ Often reaching back to the Internet Archive
- ▶ Some technical info on WIKI pages
- ▶ Science oriented blogs are appearing
- ▶ Collaborative workbenches emerging
- ▶ More instant messaging, especially next generation
- ▶ What can we and what should we capture?
- ▶ What will science historians do in 50 years?



Public Use of the SkyServer

- ▶ Prototype in data publishing
 - 500 million web hits in 6 years
 - 1,000,000 distinct users vs 15,000 astronomers
 - Delivered 50,000 hours of lectures to high schools
 - Delivered >100B rows of data
 - Everything is a power law
 - ▶ Interactive workbench
 - Casjobs/MyDB
 - Power users get their own database, no time limits
 - They can store their data server-side, link to main data
 - Simple analysis tools (plots, etc)
- Over 2,400 'power users'



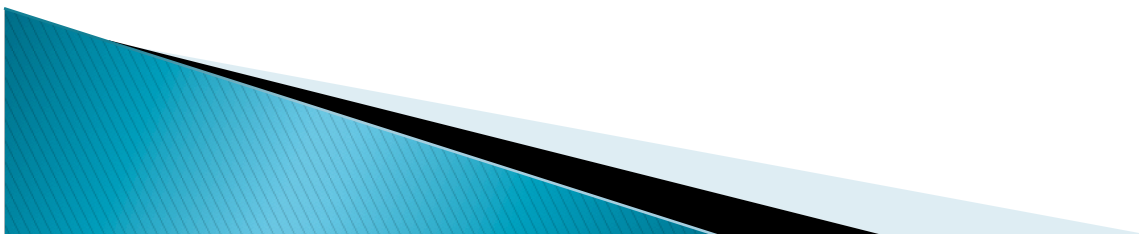
GalaxyZoo

- ▶ Built on top of SkyServer
 - ▶ 27 million visual galaxy classifications by the public
 - ▶ Enormous publicity (CNN, Times, Washington Post, BBC)
 - ▶ 100,000 people participating, blogs, poems,
 - ▶ Now truly amazing original discovery by a schoolteacher
 - ▶ Observations scheduled on Hubble, VLBA
-
- ▶ A new pattern in using scientific data!



National Virtual Observatory

- ▶ NSF ITR project, “Building the Framework for the National Virtual Observatory” is a collaboration of 17 funded and 3 unfunded organizations
 - Astronomy data centers
 - National observatories
 - Supercomputer centers
 - University departments
 - Computer science/information technology specialists
- ▶ Similar projects now in 15 countries world-wide
=> International Virtual Observatory Alliance



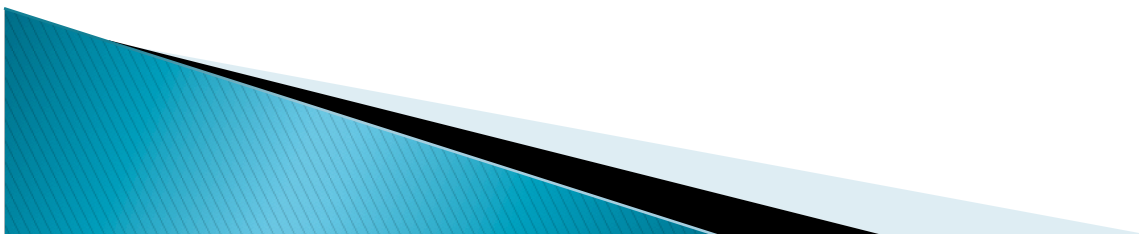
Current Status

- ▶ The project has ended on Nov 1, 2008
 - ▶ Most of our current people now part time
- ▶ The sociological transformation successfully done:
 - Main data providers now all offer compliant services
- ▶ Service-oriented architecture, before Web 2.0 wave
- ▶ Deliberately NOT a full top-down design
- ▶ Creation of standards took longer than expected
- ▶ Need to transition from research to proper facility
- ▶ NSF/NASA AO for VAO announced in 2008
- ▶ Joint proposal by AUI/AURA submitted in Apr 08
- ▶ **It is inevitable!**



Data Sharing/Publishing

- ▶ What is the business model (reward/career benefit)?
- ▶ Three tiers (power law!!!)
 - (a) big projects
 - (b) value added, refereed products
 - (c) ad-hoc data, on-line sensors, images, outreach info
- ▶ We have largely done (a), mandated by NSF/NASA
- ▶ Need “Journal for Data” to solve (b)
- ▶ Need “VO-Flickr” (a simple interface) for (c)
- ▶ Mashups are emerging (GalaxyZoo)
- ▶ New public interfaces to astro data (Google Sky, WWT)
- ▶ Integrated environment for
‘*virtual excursions*’ for education (C. Wong)

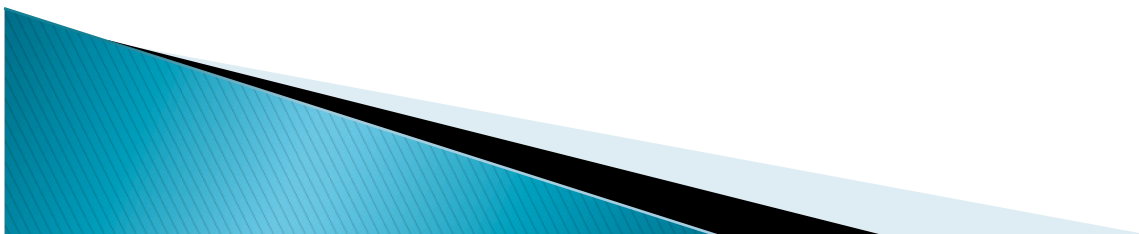


‘Journal of Data’ in Astronomy

Create new paradigm in publishing scientific data

- Team up with the main journals in astronomy
- On-line supplement for data related to journal articles
- Easy submission process for authors
- Data replicated among university libraries
- Data guaranteed to exist for 20 years
- Uses Fedora Commons
- Curation, curation, curation!!!

with S. Choudhury, T. DeLauro (JHU Eisenhower Lib), R. Hanisch (Space Telescope), E. Vishniac (McMaster), C. Lagoze (Cornell)



Continuing Growth

How long does the data growth continue?

- ▶ High end always linear
- ▶ Exponential comes from technology + economics
 - ↔ **rapidly changing generations**
 - like CCD's replacing plates, and become ever cheaper
- ▶ How many new generations of instruments do we have left?
- ▶ Are there new growth areas emerging?
- ▶ **Software is becoming a new instrument**
 - Simulations!!
 - hierarchical data replication
 - Value added data/ mashups

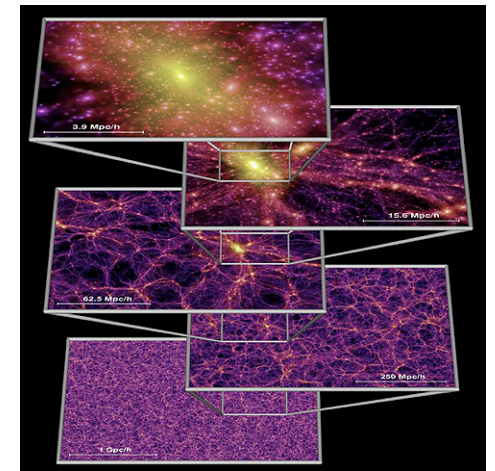


Simulations

Cosmological simulations have 10^9 particles and produce over 30TB of data (Millennium, Aquarius, VLII)

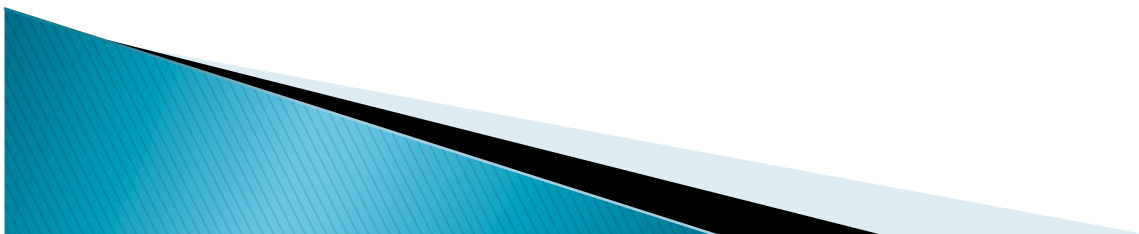
- ▶ Build up dark matter halos
- ▶ Track merging history of halos
- ▶ Use it to assign star formation history
- ▶ Combination with spectral synthesis
- ▶ Realistic distribution of galaxy types

- ▶ Too few realizations
- ▶ Hard to analyze the data afterwards → need DB (Lemson)
- ▶ What is the best way to compare to real data?
- ▶ Data volumes soon reaching Petabytes



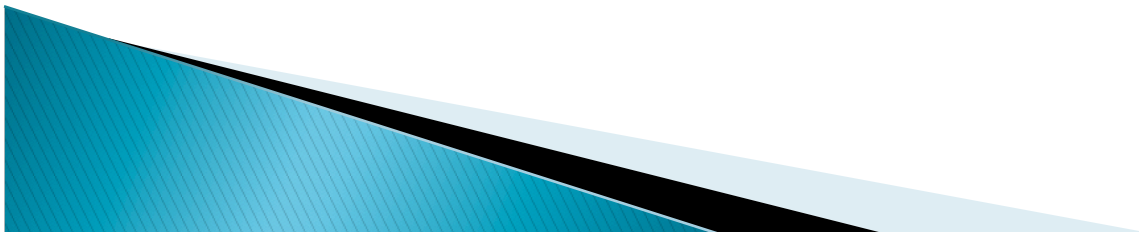
VO Technology

- ▶ The next surveys will generate Petabytes
- ▶ We will need to save them, move them
 - several big archive centers connected
 - shakeout
- ▶ Archives -- also computational services
 - driven by economics: cheaper to process than move
- ▶ Always an open-ended modular system
- ▶ Need Journal for Data
 - curation is the key



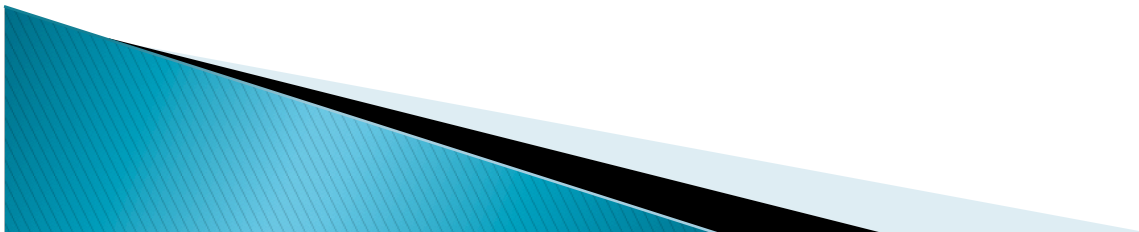
VO Economics

- ▶ The Price of Software
 - 30% from SDSS, 50% for LSST
 - should there be full reuse vs no reuse today?
 - neither: we are not systems integrators
 - risks and benefits are power law
- ▶ The Price of Data
 - \$100,000 /paper (Norris et al)
 - Drives new projects
 - For SDSS there are 3,000 refereed papers for \$100M
- ▶ Level budgets



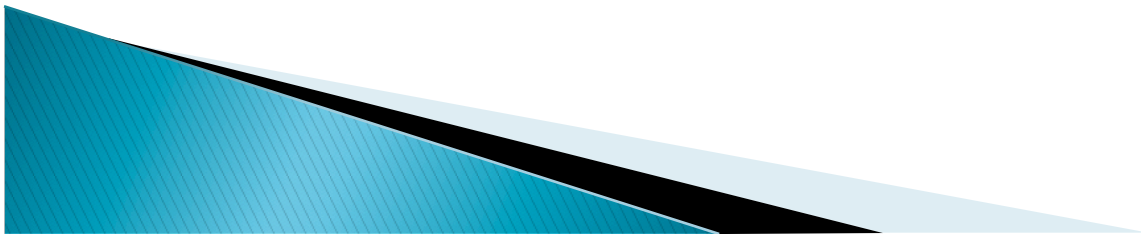
VO Sociology

- ▶ Learn from particle physics
 - do not for granted that there will be a next one
 - small is beautiful
- ▶ What happens to the rest of astronomy after the world's biggest telescope?
- ▶ The impact of power laws:
 - we need to look at problems in octaves
 - the astronomers may be the tail of our users
 - there is never a natural end or an edge (except for our funding)
- ▶ Unpredictable changes, new players



Collaborative Trends

- ▶ Science is aggregating into ever larger projects
- ▶ Collection of data **increasingly separated** from analysis, connected with the data publications
- ▶ VO is inevitable, a new way of doing science, present on every physical scale today
- ▶ Natural size for close collaborations is small
- ▶ May be the only way to do 'small science' in 2020



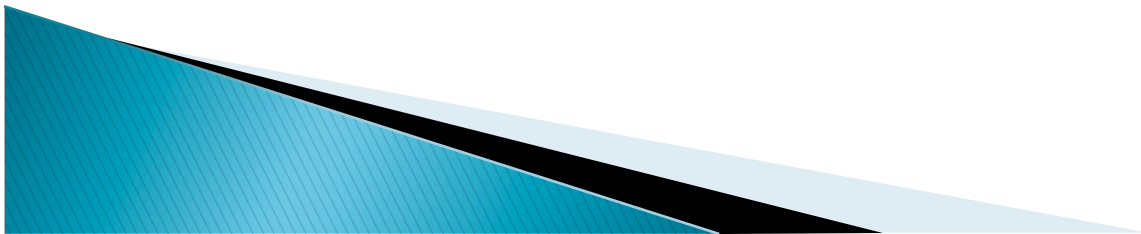
Near Future

- ▶ Surveys role increasing, more archival data
- ▶ Relatively easy to predict until 2010
 - Exponential growth continues
 - Most ground based observatories join the VO
 - More and more sky surveys in different wavebands
 - All sky Xray survey is missing, nothing since ROSAT
- ▶ Dominance of Large Imaging Surveys
 - Fastest explosion of data in radio
 - Urgently need large wide field spectroscopy survey!
- ▶ Simulations will reach petabytes
 - Will have VO interfaces: can be 'observed'



Beyond 2010

- ▶ PetaSurveys are coming on line (Pan-STARRS, VISTA, LSST) and becoming public
- ▶ Petabytes will need a hierarchical organization
 - Need a proper “impedance match”
- ▶ Single Query analysis paradigm will break
- ▶ Expect world-wide network of large archive/compute centers
- ▶ Business model unclear: public data does not necessarily mean accessible data...
- ▶ Moore’s Law comes to the rescue (up to a point)
- ▶ Changing funding climate, unpredictable



“The future is already here. It’s just not very evenly distributed”

William Gibson

